Modeling the Acoustic Correlates of Expressive Elements in Text Genres for Expressive Text-to-Speech Synthesis

Hongwu Yang[1, 2], *Helen M. Meng*[2], *Lianhong Cai*[1]

[1]Department of Computer Science and Technology
Tsinghua University, 100084 Beijing, China
[2]Department of Systems Engineering and Engineering Management
Chinese University of Hong Kong, HKSAR, China
yang-hw03@mails.tsinghua.edu.cn, hmmeng@se.cuhk.edu.hk, clh-dcs@tsinghua.edu.cn

Abstract

This paper proposes a novel approach for describing the expressive elements in text genres and modeling their acoustic correlates for expressive text-to-speech synthesis (TTS). We apply the three-dimensional PAD (pleasure-displeasure, arousal-nonarousal and dominance-submissiveness) model in describing expressivity. In particular, we define a set of principles for annotating the *P* and *A* values of prosodic words found in texts from the tourist information domain. These text passages may be categorized into the descriptive genre (e.g. describing a beautiful scenic spot), the informative genre (e.g. presenting the opening hours of a museum) and the procedural genre (e.g. offering bus routes to a landmark). We choose the prosodic word as the basic unit for analysis since it bridges textual input with (synthetic) speech output. Analysis of contrastive (neutral versus expressive) recordings uncovers the acoustic correlates of annotated *P* and *A* values. This enables us to develop a non-linear model that can transform neutral speech to resemble expressive speech, according to the *P* and *A* values of the input text. Perceptual evaluation of the speech outputs shows that over 70% of the prosodic words carry appropriate expressivity.

Index Terms: expressive speech synthesis, text-to-speech, nonlinear model, text genres, PAD emotional model

1. Introduction

Expressive speech synthesis has been a hot topic of research in recent years [1]-[4]. Primary emotions such as the “big six” [5] have been widely accepted and used in many studies [2],[6]. It has strong potential in enhancing effective communication between human and computers in spoken dialog systems. Expressivity may be a function of the speaker’s internal state, the state of the dialog, the intended effect for the listener as well as the message content. This work focuses on the message content. Our long term objective is to incorporate expressive TTS for response generation in a spoken dialog system that supports user inquiries in the Hong Kong tourism domain. We observe four main types of message genres in these computer-generated responses: (i) the *interactive genre*, where the message aims to carry forward to the next dialog turn or bring the dialog to a close; (ii) the *descriptive genre*, where the message describes the beauty or specialties of a scenic spot; (iii) the *informative genre*, where the message present facts (e.g. opening hours of a tourist spot) and (iv) the *procedural genre*, where the message gives directions (e.g. driving directions). As an initial step, this study focuses on the descriptive, informative and procedural genres, since these are primarily based on the textual content of the message, i.e. text genres. In the following, we will describe our work in finding descriptors of expressivity for text genres, finding acoustic correlates of these descriptors, modeling the correlation between the descriptors and their acoustics, as well as using the model to modulate neutral (expressionless) speech. We believe that such a model can be used to modulate the neutral outputs of our TTS systems based on the genres of the input text. In this way, we will be able to generate expressive, synthesized speech for our spoken dialog system.

<name of tourist spot>海洋公園 (English: Ocean Park)
<descriptive text>
香港海洋公園是全東南亞規模最大的水族及主題公園之一。您可以在園內參觀大型的珊瑚礁水族館，觀賞海豚表演，探訪可愛的國寶大熊貓“安安”和“佳佳”，嘗試機動城內多種緊張刺激的機動遊戲，玩個不亦樂乎。
(English: One of Southeast Asia’s largest oceanariums and theme parks, featuring aquariums, dolphin shows, thrilling rides, giant pandas An An and Jia Jia, and much more.)
<informative text>
開放時間為每日上午十點至下午六點。
(English: Open daily, 10am-6pm.)
<procedural text>
您可於地鐵金鐘站 B 出口或中環天星碼頭附近（地鐵中環站 K 出口）乘城巴 629 路往返海洋公園
(English: Special City bus 629 leaves from near the Star Ferry Piers (Central MTR Exit K) in Central and Admiralty MTR Exit B.)

Figure 1 *An example of a text passage about a tourist spot.*

2. Scope and Text Genres

The scope of the current study lies in the Hong Kong tourist information domain, which is the same as that for our spoken dialog system. We sourced content from the Hong Kong Tourism Board [7], which includes a passage for every key tourist spot. The typical format of each passage is shown in Figure 1. It begins with descriptive paragraph, followed by an informative paragraph about opening hours and/or ticket prices and finally a procedural paragraph about transportation and walking directions. This study focuses on the Chinese content, but we also include the English translations for the purpose of readability.

As can be seen, the descriptive text often contains commendatory words that describe scenic characteristics or specialties about the location. Synthesizing a spoken presentation of such descriptions should incorporate the appropriate prosody. The informative text and procedural text contain useful facts for the tourist. Synthesizing a spoken presentation of these facts should incorporate appropriate emphasis to draw attention from the listener.

We have included the passages corresponding to twenty popular tourist spots in this study. Hence our text corpus contains 60 paragraphs, which corresponds to 1,358 Chinese prosodic words and 3,340 syllables in total. We have chosen the prosodic word as the basic unit for analysis since it bridges the textual input with the synthetic speech output. The average syllable count per prosodic word is 2.5.

3. The PAD Model

Previous work has used methods such as categorical annotation scheme [8] and two-dimensional annotation scheme [9] to annotate emotions. In this work, we seek to use a compact set of descriptors that can capture significant variability in expressivity across possible dialog responses in our domain. We adopted the PAD model proposed by Mehrabian [10] as the descriptor to annotate the expressivity (or sentiment) of a prosodic word from text. The PAD model describes and measures emotional states along three nearly independent dimensions: "Pleasure-displeasure" (P) distinguishes the positive-negative affective quality of emotional states, "arousal-nonarousal" (A) refers to a combination of physical activity and mental alertness, and "dominance-submissiveness" (D) is defined in terms of control versus lack of control. The implementation of the PAD three-dimensional space uses axes ranging from -1.0 to 1.0 for each dimension.

4. Annotating Expressivity in Text

The textual passages (see figure 1) are automatically segmented into prosodic words by means of a home-grown software tool that applies a set of heuristic rules previously induced from data. We laid down a set of principles for annotating the expressivity of prosodic words.

4.1. Principles of Annotation

(i) **P values:** Commendatory words or words with positive connotations in the descriptive text are labeled with *P=1* (for pleasure). Derogatory words or words with negative connotations are labeled with *P=-1* (but these have not been observed due to the characteristics of our domain). Other words are labeled with *P=0*.

(ii) **A values:** Superlatives and words denoting a high level or extent are labeled with a maximum degree of arousal, i.e. *A=1*. Comparatives and words carrying key facts which should be emphasized (e.g. street names, transportation means, etc.) are labeled with an intermediate degree of arousal, i.e. *A=0.5*. Other words are labeled with *A=0*. A common construct found in our current corpus is "…not only <phrase1>, but also <phrase2>…." We annotate prosodic words in <phrase1> with *A=0.5* and those in <phrase2> with *A=1*.

(iii) **D values:** We consider that annotation for dominance-submissiveness (*D*) are not relevant to the current text corpus but will become relevant when we include the interactive genre of textual responses.

4.2. Results of Annotation

Three annotators were invited to annotate the text corpus and to follow the principles as closely as possible. The final *P* and *A* values adopted for each prosodic word is determined based on the principle of majority. Table 1 shows statistics related to the annotated prosodic words. It can be seen that most prosodic words are labeled with neutral *P=0* values (85.4%). Over half (67.5%) of the prosodic words are labeled with (*P=0, A=0.5*), mainly because both the informative and procedural text genres have common occurrences of prosodic words presenting facts that should be emphatically synthesized. No prosodic word is found to carry the annotations of (*P=0 and A=1*) or (*P=1 and A=0*), which implies that a word that is neutral on the pleasure-displeasure dimension will not carry a high extent of arousal. Conversely, a word that indicates pleasure (i.e. a commendatory or descriptive word) must carry some degree of arousal. With this observation, our study proceeded with focus on the four types of (*P,A*) combinations, namely (*P=0 and A=0*), (*P=0 and A=0.5*), (*P=1 and A=0.5*), (*P=1 and A=1*). Table 2 shows annotated prosodic words in phrase about Repulse Bay.

Table 1. *Statistics of annotated P and A values for prosodic words (PW) in our tourist information text corpus.*

(P,A)	(0,0)	(0,0.5)	(0,1)	(1,0)	(1,0.5)	(1,1)
#PW (total: 1358)	238	922	0	0	141	57
% of occurrence	17.5	67.9	0	0	10.4	4.2

Table 2. *Examples of annotated prosodic words (PW) in a phrase about Repulse Bay. Meaning of the tabulated Chinese words (from left to right) are: "this", "crescent-shaped", "beach", "is Hong Kong's", "most", "popular", "beaches".*

pw	這個	呈半月型的	沙灘	是香港	最受	歡迎的	海灘之一
(P,A)	(0,0)	(0,0.5)	(0,0.5)	(0,0.5)	(1,1)	(1,1)	(0,0.5)

5. Acoustic Realization of Expressivity

5.1. Speech Recordings

We proceeded to investigate how expressivity relating to prosodic words from text is realized acoustically. The acoustic analysis is based on a speech corpus especially designed with contrastive speech recordings (neutral versus expressive) for each of the 60 paragraphs in the twenty passages of our text corpus. A native Cantonese female speaker was invited to record in a studio. There are (20passages)*(3paragraphs)*(2neutral/expressive)=120 speech files in total, amounting to about 45 minutes of speech. All recordings were saved in the Microsoft Windows WAV format as sound files (mono-channel, unsigned 16 bit, sampled at 16kHz).

5.2. Acoustic features

Our objective is to analyze how expressive elements from text may be realized in the acoustic speech signal. Acoustic features that are commonly associated with prosody include fundamental frequency (F0), intensity, speaking rate and pause durations. Therefore we choose to focus on the acoustic measurements F0, the speaking rate, fluency and intensity. We capture both the average and the dynamicity of these acoustic features in seven measurements from each prosodic word:

- **Intonation:** mean, range, slope;
- **Intensity:** mean and range of the RMS energy (rmsmean and rmsrange);
- **Fluency:** duration of pause between the prosodic word;
- **Speaking rate:** syllables per minute.

5.3. Acoustic correlates of expressive features

The recorded utterances are first automatically segmented into syllables with a home-grown segmentation tool and then the syllable boundaries are checked manually. Measurements are taken from the contrastive recordings (neutral versus expressive) of each paragraph (which may be descriptive, informative or procedural). We also compute the percentage increase in the values of the measurements as one migrates from the neutral speech to its expressive counterpart. Results are shown in Figure 2 for different combinations of *(P, A)* values. The *A* values are shown on the x-axis, triangles denote cases when *(P=0)* and circles denote cases when *(P=1)*.

6. Nonlinear model for acoustic correlates of expressive features

Based on the observations of Figure 2, we propose a nonlinear model to capture the relationship between the *P* and *A* values with the acoustic measurements, as shown in Equation 1.

$$\frac{F^{\exp}}{F^{neu}} = C_1 P \exp\left(-C_2 A\right) + C_3 A \exp\left(-C_4 P\right) + C_5 \qquad (1)$$

where $F^{\exp}$ is the feature from expressive speech, F^{neu} is the feature from neutral speech, $F^{\exp} / F^{neu}$ is the increase of the expressive feature from the neutral feature, and $C_1 \cdots C_5$ are constants. This equation captures that when (*P=0*), the relative increase of the expressive measurement bears a linear relationship with that of the neutral measurement, as we interpolate an increase of *A* from *0 to 0.5*. This relationship is changed when *(P=1)* by a factor of $C_1 exp(-C_2 A)$, as we interpolate an increase of *A* from *0.5 to 1*. We use nonlinear least-squares regression to estimate the constant values in Equation (1). In order to produce an accurate finite-difference gradient, the initial coefficients are set to 1, the maximum number of iterations is 100. Results of regression are shown in Figure 3.

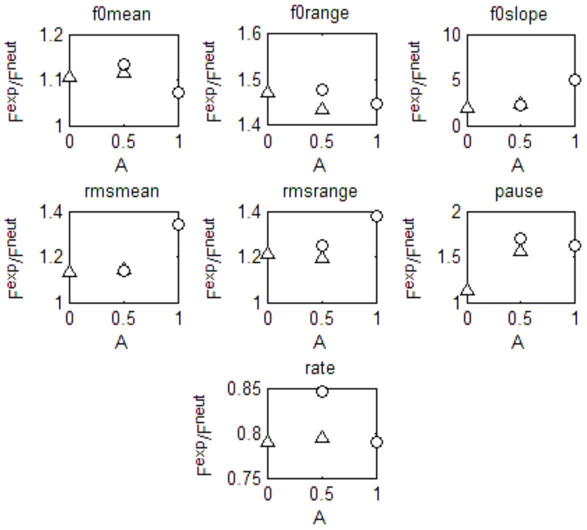

Figure 2 *Difference of expressive features from neutral features for different P and A.*

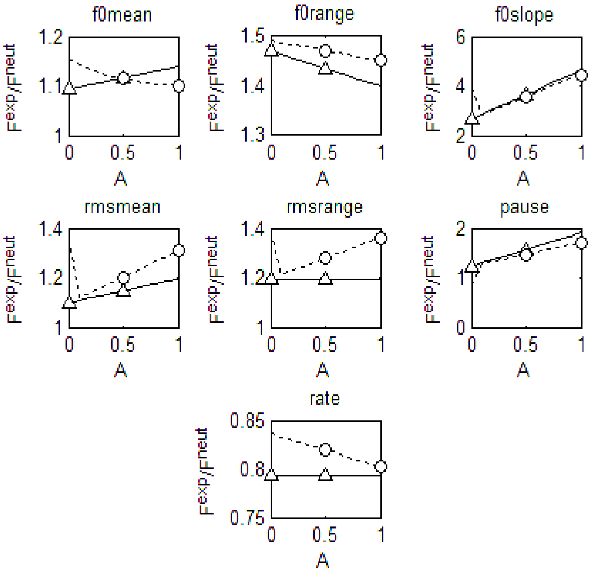

Figure 3 *Regression results for the non-linear model.*

7. Evaluation of the Nonlinear Model

We devised a set of preliminary experiments to evaluate the nonlinear model. We selected twenty textual sentences within our tourist information domain and tokenize them into prosodic words using the homegrown software tool. We also ensure that the annotated prosodic words within this set have a good coverage of the distributions in the *P-A* space. (The annotations are based on the principles laid out in section 4.) We organize a perceptual evaluation whereby each textual sentence is presented to a subject as three speech audio files: (I) a speech recording of neutral speech from the same female speaker mentioned above; (II) a speech recording of expressive speech from the same speaker; and (III) a *transformation* of the speech file from the neutral recording. In this transformation, we use the annotated *P* and *A* values of each prosodic word to obtain F^{exp}/F^{neu} values based on the nonlinear model described in Equation 1. These parameters are used to transform the speech segment of the corresponding prosodic word in the recorded neutral waveform. Six of acoustic measurements (all except for pause duration) are modified by the use of STRAIGHT [12]. The pause duration is concatenated to the ends of the prosodic word in a subsequent step. Transformed speech segments from all the prosodic words are concatenated in order to form the modified speech utterance with synthetic expressivity, i.e. (III). We invited 12 native speakers of Cantonese to be our subjects in a listening evaluation. For each of the textual sentences, the speech files are played for the subjects in the order of I-II-III-I-II-III. While listening, the subject sees a listing of all the prosodic words in the sentence and judges whether a prosodic word in (III) more closely resembles its counterpart in (I) versus that in (II). Results are shown in Table 4.

Table 4. *Results of the listening evaluation. % PW denotes the percentage of transformed prosodic words judged to bear closer resemblance with its counterpart in (II) (expressive version) versus that in (I) (neutral version).*

(P, A) values	(0,0)	(0,0.5)	(1,0.5)	(1,1)
# prosodic words (PW)	17	76	14	15
% PW	70.6	73.2	84.5	76.1

8. Conclusions and Future Work

This paper proposes a novel approach for describing the expressive elements in text genres and modeling their acoustic correlates for expressive text-to-speech synthesis (TTS). We apply the three-dimensional PAD (pleasure-displeasure, arousal-nonarousal and dominance-submissiveness) model in describing expressivity. In particular, we found that only the *P* and *A* values are directly applicable in textual content sourced from the tourist information domain. These text passages may be categorized into the *descriptive, informative* and *procedural* genres. We tokenized the text into prosodic words using a homegrown software tool. We choose to focus on the prosodic word because the unit can bridge text input with (synthetic) speech output. We developed a set of principles of annotation and manually labeled prosodic words with *P* and *A* values. Analysis of contrastive (neutral versus expressive) recordings uncovers the acoustic correlates of annotated *P* and *A* values. This enables us to develop a non-linear model that can transform neutral speech to become expressive speech, according to the *P* and *A* values of the input text. Perceptual evaluation of the speech outputs shows that over 76% of the prosodic words carry appropriate expressivity. Future work will attempt to extend the nonlinear model to cover the *interactive text genre* (mainly from computer-generated response text in a spoken dialog system). The nonlinear model will then be used to enhance the expressivity of our existing Chinese text-to-speech synthesizers for response generation in a spoken dialog system for the tourist information domain.

9. Acknowledgements

This work is supported by the research fund from the National Science Foundation of China (NSFC) under grant No. 60433030 and the joint fund of NSFC-RGC (Research Grant Council of Hong Kong) under grant No. 60418012 and N-CUHK417/04. This work is affiliated with the Microsoft-CUHK Joint Laboratory for Human-centric Computing and Interface Technologies.

10. References

[1] Campbell, N., “Towards Synthesizing Expressive Speech: Designing and Collecting Expressive Speech Data,” Proc. Eurospeech 2003, pp. 1637-1640, 2003.

[2] Hamza, W., Bakis, R., Eide, E., et al., “The IBM expressive speech synthesis system,” Proc. ICSLP 2004.

[3] Bulut, M., Narayanan, S. and Johnson, J., “Synthesizing expressive speech: overview, challenges, and Open Questions,” In S. Narayanan and A. Alwan. Text-to-Speech Synthesis: New Paradigms and Advances, pp. 175-201, Prentice Hall.

[4] Tsuzuki, R., Zen, H., Tokuda, K., Kitamura, T., Bulut, M. and Narayanan. S., "Constructing Emotional Speech Synthesizers with Limited Speech Database.” Proc. ICSLP 2004.

[5] Cowie, R., Douglas-Cowie, E., Tsapatsoulis, N., et al., “Emotion recognition in human-computer interaction,” IEEE Signal Processing Magazine, 2001.

[6] Eide, E., Bakis, R., Hamza, W. and Pitrelli, J. F., “Toward expressive synthetic speech,”, In Narayanan, S. and Alwan, A., Text-to-Speech Synthesis: New Paradigms and Advances, pp. 219-248, Prentice Hall.

[7] http://www.discoverhongkong.com.

[8] Martin, J. C., Pelachaud, C., Abrilian, S., Devillers, L., Lamolle, M. and Mancini, M., “Levels of Representation in the Annotation of Emotion for the Specification of Expressivity in ECAs,” IVA’05 International Working Conference on Intelligent.

[9] Craggs, R., “Annotating emotion in dialog: issues and approaches,” In Lee, M. (Ed), Proceedings of the 7th Annual CLUK Research Colloquium, 2004.

[10] Mehrabian, A., “Correlations of the PAD Emotion Scales with self-reported satisfaction in marriage and work,” Genet Soc Gen Psychol Monogr. 124(3):311-34, 1998.

[11] Lee, C. M., Narayanan, S. and Pieraccini, R., “Combining acoustic and language information for emotion recognition”, Proc. ICSLP 2002.

[12] Kawahara, H., Estill, J. and Fujimura, O., “Aperiodicity extraction and control using mixed mode excitation and group delay manipulation for a high quality speech analysis, modification and synthesis system STRAIGHT,” MAVEBA 2001, Firentze Italy, 2001.